Data Management Process Example Narrative

March 2018

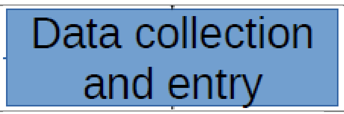
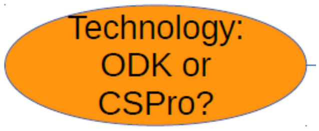
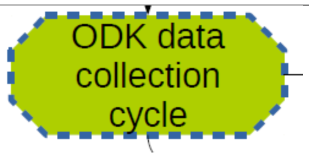
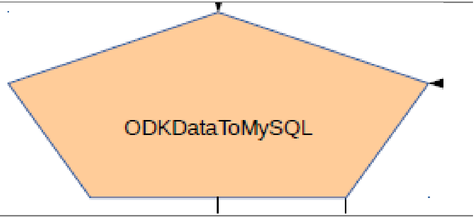
Table of Contents

Introduction

This document aims to provide some narrative to the flowcharts of the ILRI Data Management Process provided in a separate PDF document. The flowchart document was produced in 2014 and, although there have been some advances in data management practices since then, the charts are still very useful in reminding researchers of the various tasks involved in the whole data management process.

The flowcharts consider both digital data collection using ODK on hand-held devices, and data collection using paper questionnaires with the subsequent data entry using CSPro.

The elements of the flowcharts are colour-coded as follows:

Blue rectangles represent steps in the process	
Orange ovals contain questions that the PI and the project team should consider	
Green shapes with dotted blue edges link to later charts on the pages that follow	
Pale orange pentagons are software tools	

ILRI's Data Management Process

Looking at the flowchart of the overall process we start at the top with the **Study Design** step and finish with archiving the data to a Data Portal in line with the CGIAR Open Access policy.

Study Design

The study design stage includes setting the objectives of the study and working out what data needs to be collected in order to fulfil those objectives. In this process you will be developing the questionnaire to be used. Questions at this stage include deciding on what standard indicators are going to be used and whether or not translations are needed. Where possible it is ideal to have the questionnaire in the local languages and to do back-translations to ensure the original meaning of the questions is not lost in translation. This is far preferable to on the spot translations by the enumerators during the interviews

Metadata

Now is the ideal time to start pulling together the metadata that will be needed when the data are archived; if you leave this task until the end then those with the necessary information are likely to have moved on to other projects and gathering the necessary information becomes very difficult and time-consuming. Think about where the data are to be published and pull together the necessary information. Remember this is the project level metadata which could be referred to as the Study Catalogue; later on, there will be the description of the dataset itself which will also need to be included – we refer to this as the Data Dictionary. See the guide “Introduction to Metadata” for further information about the levels of metadata.

Data Collection and Entry

We then move onto the process of collecting and entering the data. The key question at this stage is about the technology to use: are you going to use digital data collection using ODK or are you going to use paper questionnaires and enter the data using CSPro? Of course, it is entirely feasible that you use digital data collection in some sites and paper questionnaires in others.

Coding the Study

Once you have decided which technology to use the next task is to create the ODK system or the CSPro system (or both). In theory data collection using paper questionnaires could be started before the CSPro system is prepared; however, it is a good idea to have the system working before data collection starts so that data entry can start as soon as some completed questionnaires become available. Of course, it is essential to have the ODK system ready before fieldwork can start.

Testing

Both the ODK system and the CSPro system must be thoroughly tested. Ideally you will be carrying out a pilot study and a pilot is the perfect opportunity for road-testing the systems. Remember both systems should include automatic checks on the data and we would recommend compiling a comprehensive list of checks to incorporate into the system. These checks might include checking that particular values are within a specified range or checking for consistency across variables. For example, in a household roster you would want to avoid situations where a child appears to be older than his/her parents. Some checks might be “hard” checks while others might be “soft” checks. For example, you might expect a mother to be at least 15 years older than her child but accept that some mothers might be as young as 10 years of age; you could therefore have a “hard” check whereby the system will not accept a mother being only 9 years older than her child and a “soft” check where you ask for confirmation if the age difference is between 10 and 15 years.

DB Creation on Server

Once you know all the variables you will be creating you can create the database on the server to hold the collation data. We will expand on the database creation process later.

Training

In the flowchart, training appears to follow the data collection. However, training should come before data collection. Regardless of whether you are using ODK or CSPro the enumerators will need to be trained to carry out the fieldwork. They will need to become familiar with the questionnaire and the types of responses that are expected – for example which questions only accept a single response and which are multiple response, which require a date, and which are open text questions. If you are using ODK, the enumerators will also need to be trained on using the ODK system. Separate training for the supervisors would cover tasks such as uploading data to the aggregate server, etc. If you are only using paper questionnaires, the enumerators will not need to know how to use the CSPro system. Instead, you will need to train data entry staff in the use of this system.

ODK & CSPro Data Collection Cycle

After the training we have the data collection phase. Data collection in ODK and in CSPro will be expanded on in the next main section.

Data in Database Server

Towards the end of the data collection cycle, the data will be collated into the database created earlier on the server.

Data Cleaning

There is then the process of data cleaning and the main question here is "Who cleans the data?". It must be remembered though, that data cleaning is not just a single phase; errors can appear in the data at various points in the project life-cycle, so it is important to remain vigilant. We would also recommend keeping an audit trail of changes and corrections made to the data. The audit trail should be included with the data archive so that you can defend your results should anyone question the validity of your data.

Enabling Audit, Access & Data Interoperability

These are all tasks related to the Research Methods Group (RMG) and Data Management Community of Practice (DCoP) at ILRI and is about allowing these groups access to the database so that they can audit the data and integrate it into their systems.

Restricted Access to Data in Data Portal

Set out in the policy documents for the organisation, there will be mention of the length of time researchers on a project can have restricted access to the data for them to be able to work on publishing their results. This is standard practice and is generally something like 12 months from data collection. At this stage you will need to determine who should have access and who is going to publish the data.

ODK Data Collection Cycle

Upload ODK to Aggregate Server

Once the study has been coded in ODK the next step is to upload it to the aggregate server. In the flowchart "Formhub" is mentioned; Formhub was an aggregate server that was popular a few years ago but there are now many others such as ONA.

Download ODK to device

Once the ODK system is on the aggregate server you need to download it to the hand-held device ready for the fieldwork. Of course, you must already have installed ODK Collect onto the device and will need to have adjusted the settings to link to the aggregate server that you are using. We would suggest that only the supervisors download the system to the devices to ensure all enumerators are using the correct version.

Testing Data Collection

Once the system has been downloaded it will need to be checked on the device. If problems are found or changes are needed then you will need to go back to the coding stage, upload the revised form and download it again to the device. It is important that all enumerators have the same working version of the ODK forms on their devices.

Data Collection by Enumerator

This is where the enumerators interview household members and collect the data. The number of households that can be visited in one day will depend on the length of the questionnaire and this will have been determined during the training.

Daily Review by Enumerator

At the end of each day there should be time for the enumerators to review the data they have collected and check for any invalid or inconsistent data. Many of these are likely to have been found by the ODK system itself but it is not always possible to check for every eventuality, so a review is still necessary.

Data Correction by Enumerator

If invalid data is found, then the enumerator will need to correct this.

Visit Household for Corrections

It may be the case that the enumerator can correct the data without revisiting the household. For example, he/she may notice a spelling mistake in the name of the village. However, there may be errors or inconsistencies that require a revisit to the household to confirm or correct values.

Copy Data to Supervisor

Whether or not corrections were needed after the enumerator review, only valid or corrected data should be passed to the supervisor. You will need to decide how many supervisors you need but of course this should be decided before the start of the fieldwork.

Supervisor Review of Data

Once the supervisor has the data from the enumerators, he/she should do his/her own review. If invalid data values are identified, then these should be returned to the relevant enumerator for correction. It may seem excessive for both the enumerator and the supervisor to review the data, but as we said earlier on, the process of data cleaning and checking is an on-going process and errors can be found at any stage.

Upload Data to Aggregate Server

Once the supervisor is satisfied that the data are valid, he/she can then upload the data to the aggregate server. We suggest that only supervisors upload data as this helps to ensure the validity of the data.

Zip and Send Data to ILRI

An alternative to uploading the data to the aggregate server is to zip the data and send it directly to ILRI to be processed. This is also after the supervisor review and any subsequent corrections.

Process ODK data from Aggregate Server or ZIP file

In a separate section we look at the stage of processing the data either from the ZIP file or from the aggregate server.

Data in Database on Server

Once the data have been processed it will be collated into the database on the server that was created earlier.

Data Analysis by Users

At this stage the researchers can start on the analysis of the data. During this process they may identify invalid data.

Invalid Data Identified

If invalid data values are identified during the analysis, then data queries should be sent back to the supervisor who will need to review the problems and arrange correction if possible. Of course, by this stage it might not be possible to revisit households to check on data, so a decision will need to be made about any errors and inconsistencies. It is important to document these decisions.

CSPro Data Collection Cycle

For data collection using paper questionnaires and subsequent data entry in CSPro, the process of coding the study in CSPro (i.e. preparing the data entry system) can be done in parallel with the fieldwork. However, we would suggest you aim to have the system completed and tested before any completed questionnaires start coming in so that data entry can be started straight away.

Coding the Study in CSPro

This step includes thoroughly testing the system. As mentioned earlier, many data checks and question skips can be automatically programmed into the system, and all such skips and checks need to be tested.

Upload CSPro onto Data Entry PCs

The final CSPro system will need to be placed on all PCs being used for data entry. You would need the CSPro software installed on the PCs and, from the data entry system itself, you would need the dictionary file (extension .dcf) and the compiled binary version of the system (extension .pen or .enc for earlier versions of CSPro).

Data Collection on Paper

Meanwhile data collection can be taking place. It is essential that the enumerators have been trained to carry out the interviews and know what is expected of them in terms of completing the questionnaires. Everything entered on the questionnaire must be clear; any errors should be crossed through neatly, so it is obvious that this is a mistake. We recommend having a section on the front of the questionnaire with space for enumerators, supervisors, and data entry staff to sign and enter the date. In signing the questionnaire, they are taking responsibility for having completed their tasks.

Review by Supervisor (paper questionnaires)

Before the completed questionnaires are sent for data entry, the supervisor should visually review the data looking out for any inconsistencies or other errors. It is useful to have a checklist for this process so that nothing is missed. You will need to decide on the number of supervisors you have. This is likely to depend on the size of the team. Note that the data entry system will catch some inconsistencies, but it is important to check the completed questionnaires while still in the field so that there is the possibility of returning to the household to double-check the data.

Data Correction by Enumerator

Where errors are found the corresponding questionnaires are returned to the relevant enumerators for correction and checking. Once corrected the questionnaires will again go to the supervisor for review.

Visit Household for Corrections

For some errors or possible inconsistencies, a return to the household is desirable. Corrections should be clearly marked on the questionnaires and once again passed to the supervisor for review.

Data Entry of Paper Survey

Data entry should be done as soon as possible after data collection. There is then a greater chance of being able to correct any errors that might be found during data entry.

When you are thinking about the data entry you will need to decide on the number of data entry clerks you are likely to need. Remember the data entry staff will need to be trained in the use of the CSPro data entry system. You also need to decide whether or not you will be using Double Data Entry (DDE). This is where two different data entry clerks enter the same data into two different files. The files are then compared, and any differences are checked against the paper questionnaires. CSPro includes a Data Compare tool to facilitate the data comparison.

Review by Supervisor (data in CSPro)

Once the data have been entered they should be reviewed again by the supervisor and corrections made where possible.

Zip and Send Data to ILRI

Valid data are then zipped and sent to ILRI for processing.

Process CSPro Data

Tools within CSPro are then used to process the data. This process is expanded in a separate section.

Error Log File

If errors are found in the data, then a log file is produced which is emailed back to the supervisor for review or back to the enumerator for correction.

Data in Database on Server

The step of processing the CSPro data leads to the data being stored in the database that was created earlier on the server.

Data Analysis by Users

At this stage the data should be ready for the researchers to start some analysis using whatever statistical software they feel is appropriate.

Invalid Data Identified

During the analysis it is possible that further errors or inconsistencies in the data will come to light. If this is the case, then the supervisor should be contacted, and corrections made. If it is not possible to correct or check the data at this stage, then a decision needs to be made about whether to include these particular data values in the analysis or to treat them as missing values. Whatever decision is made it should be documented along with the reasons for that decision.

The remaining flowcharts in the accompanying file give details about the processes at ILRI. They are useful here as examples of how the data might be processed and below we briefly summarise each of these steps.

Process ODK Data using Zip Files

When the data are sent to ILRI in Zip Files it is expected to arrive in XML format. The ODK form itself is also in XML format and both the data and the form are fed into a software tool which converts ODK data to MySQL. The two outputs from the software tool are (i) the data in the database on the server, and (ii) an error log. The error log, as reported earlier, would go back to the supervisor for review and data corrections.

Process ODK Data using Formhub

The process here is very similar to the previous process. In this process there are two software tools: the first converts to the data from Formhub into JSON which the second converts JSON to MySQL. The second software tool also takes as input the ODK form in XML format. Here again the result is data in the database on the server and an error log.

Process CSPro Data

For the CSPro data, the dictionary (stored in the .dcf file) is converted to XML format. Meanwhile the data is taken from the zip file as an ASCII file (with the extension .dat) and, together with the XML version of the dictionary, is fed into a software tool that converts CSPro data to SQLite. The XML version of the dictionary file is converted to Python Script, and this, together with the SQLite version of the data, goes through a Python import script which stores the data in the database on the server and produces an error log which goes back to the data entry team.

Database Creation on Server

The database itself is created by first converting the ODK form and the CSPro dictionary to MySQL. Various scripts are run to add in the metadata and unique identifiers and basically prepare the database to be populated when the data have been collected and checked.

Summary

As we have mentioned, the database creation and the processing of the data to transfer it to the database are particular to ILRI but the flowcharts give an idea of the sort of processes that are required.

The overall data management process and the data collection cycles in both ODK and CSPro are nicely detailed and in this narrative, we have tried to include additional comments and things that should be considered at each stage. The processes include a lot of data checking and reviewing, and some might be tempted to think that this is excessive. However, as we have mentioned, errors can creep into the data at any stage, so it is important to remain vigilant so that you have a dataset that is as free from errors as possible for your analysis.